Project Title:

Exploring the opportunities and challenges of implementing open research strategies within development institutions

IDRC Project 108131
University of Cape Town
South Africa
Lynn Woolfrey, DataFirst, University of Cape Town
Final Technical Report
25 January 2017

Executive Summary

The IDRC project *Exploring the opportunities and challenges of implementing open research strategies within development institutions* was designed to inform any Open Data policies of development research funders. For this purpose, project investigators selected existing IDRC projects as Open Data case studies. Our project, the *Data on Alcohol and Tobacco in Africa* (*DATA) Project* was selected as one of the case studies. Our project aims to collect and publish tobacco-related data from selected African countries, to benefit tobacco-focused policy research. This will help African governments plan for tobacco control in their countries, to support public health objectives. At an IDRC workshop in March 2016 we were introduced to the aims of the IDRC project, and discussed our contributions with project teams from other grantee projects selected as case studies. Projects were assisted to create inventories of the data they had collected, or will collect. At the workshop, we also initiated a Research Data Management Plan for our project, which serves as a guide for how our project will manage and share the data we are collecting. Ongoing support for the IDRC project included attending the 2016 SciDataCon conference and presenting on our case study.

The IDRC project hoped to gain insight on Open Data challenges from the case studies. Our main challenge was obtaining permissions to share the data more widely. Ownership issues meant we could not share tobacco-related data from international surveys of UN organisations. Therefore, to aid data discovery, we have published information on and links to this data from our project site. We were also unable to share data from African National Statistics Agencies. We have provided links to the data on these sites, but technical constraints may mean researchers may not be able to obtain these data. Finally, the data held by the DATA Project's Principle Investigators could not always be made open. This is because data usage permissions obtained by these PIs were strictly for their projects. This information is useful for the IDRC's investigation, however, as these obstacles are common in developing countries.

We suggest the IDRC encourage existing projects to collate existing data on their topics from their countries and make these data more widely available, as researchers are being charged for data that is in the public domain, but difficult to source and use. This could benefit data-intensive research in these countries.

Research Problem

The objective of Project 108131 was the creation of a model research data policy for development research funders. The idea is that such a policy would assist these funders to support better access to development data produced by their projects. Currently much of this data is not open for on-going analysis. Specifically, the project sought to draw up Open Data guideline for funders, investigate obstacles to data sharing, and develop the data management capacities of IDRC grantees. The strategy adopted to inform the guidelines was the examination of IDRC grantee projects as case studies.

Our Project, the *Data on Alcohol and Tobacco in Africa* (*DATA) Project*, (Project 108098) was selected by the IDRC Project 108131 as one of the Open Data pilot studies. Our project aims to collect and publish tobacco-related data from selected African countries, to build a clearinghouse of data for policy analysis. Such a clearinghouse can benefit tobacco-focused research which will help African governments plan for tobacco control in their countries, to support public health objectives. In

year one of the DATA project, we focused on collecting and disseminating tobacco-related data from selected African countries, and this will be the Open Data pilot study.

The DATA Project is motivated by the paucity of access to tobacco-related data on African countries, despite the urgency for African governments to create informed tobacco-control policies to protect the health of their populations. In 2016, the World Health Organisation (WHO) estimated that tobacco use kills around 6 million people each year (WHO fact sheet, 2016). The WHO's 2003 Framework Convention on Tobacco Control (FCTC) is a guide for governments on measures to reduce tobacco use and exposure in their countries to eliminate tobacco-related morbidity and mortality. It is also an agreement among governments to do so. The treaty highlights the need for sound country data as an evidence base for effective tobacco control policies. For example, it requires parties to the treaty to:

- Integrate tobacco surveillance programmes with local and international health surveillance programmes to ensure data comparability
- Work with the WHO on standards for collecting, analysing, and publishing tobacco-related surveillance data
- Build and update databases of country tobacco control legislation
- Preserve and update data from national surveillance programmes (WHO, 2003:19)

Many of the research findings related to the economics of tobacco control are consistent across countries. For example, the finding that comprehensive advertising bans are substantially more effective than partial bans. However, policymakers typically want to know that the evidence applies to their countries as well. Country-specific research using local data thus plays a vital role in supporting public policies to control tobacco use. In some countries, such an evidence-base has led to tobacco control policies that have curbed tobacco use. This evidence was first built in high-income countries. However, the past two decades have seen a substantial volume of research emanating from low- and middle-income countries (LMICs). This is an important development, as nearly 80% of the world's 1 billion smokers live in LMICs (WHO fact sheet, 2016)

However, lack of access to this detailed country data frustrates tobacco control researchers in LMICs. Our Data on Alcohol and Tobacco in Africa (DATA) Project aims to advertise, collect, preserve, and publish African tobacco data (in year one) and alcohol data (in year two), with a focus on economic data, and on taxation and prices. This will enable African policy analysts to back up effective public policies on tobacco control with sound analyses that use solid data. Research funders are sponsoring academics in South Africa and other African countries to undertake research on tobacco control policymaking in Africa, focusing on economic issues. Building an African tobacco data clearinghouse has added another dimension to this support. As technical partners on this project, DataFirst has experience in confronting obstacles to opening African data. We have faced several of these challenges during the pilot study, and these have informed the IDRCs investigation.

Project Milestones

Project milestones for our Open Data pilot component of the project included:

- The project lead attending an introductory workshop at the IDRC in March 2016
- Drawing up an inventory of data to be collected by the project
- Drawing up and implementing a Research Data Management Plan for our project
- Assisting IDRC researchers to monitor and report on our RDM planning
- Providing guidance for further development of open research data policy implementation guidelines of development research funders

- Participating in a wrap-up meeting at the IDRC at the end of 2016

IDRC Workshop March 2016

In March 2016, we attended the introductory workshop with other grantees whose projects were selected as pilot Open Data case studies. Grantees shared the background and motivation for their projects and project aims, and elaborated on what motivated them to participate in a programme focused on data sharing. DataFirst's mandate is to share research data for further research. We could therefore share our experience with other participants who had not worked with open data. Attendees were collecting a variety of data types, and gave us insight into the data preparation and data protection requirements of these data.

Data Inventory

At the workshop, we drew up an inventory of targeted data sources. A data inventory is an excellent first step for any data-focused project. Efficient discovery tools are vital for data access, and, in our case, there is a dearth of information on what tobacco data exists in Africa. Our inventory is a working document, and we revised it after further consultation with the Project PI and Project Manager. Included in this list are data on tobacco farming, tobacco trade, tobacco products, tobacco use (prevalence, expenditure, initiation, and cessation), tobacco taxes, revenue, costs of tobacco, and the legal environment around tobacco. Our data inventory is *Appendix A* of this report. Our inventory included information on data sources and data types, as well as topic coverage and anticipated problems.

1. Data Sources include:

- International organisations, including the World Health Organisation (WHO) and other UN bodies
- African National Statistics Agencies, via data portals set up by the World Bank's International Household Survey Network. DataFirst has been involved in this project.
- Other government departments, in the form of transactional data
- Tobacco producers and industry bodies in project countries
- Principle Researchers on the DATA project

2. Data Types

Time-series Data: Governments collect tobacco economic data as part of routine administration procedures. Administrative records are continuously collected and therefore can provide tobacco-related time series. These types of data are valuable for policy analysis. For example, in modelling the likely impact of an excise tax change on government revenue or tobacco consumption, researchers are more dependent on time-series data than on cross-sectional data. In addition, researchers who want to illustrate trends in certain variables over time are dependent on time-series data. We aim to source such data and seek to obtain permissions to collate and share these data. Tobacco producers and industry bodies also publish transactional data that are useful for tobacco control research.

Cross-sectional survey and panel survey data: These data allow researchers to ask very specific questions, which are in most cases impossible to answer with time-series data. For example, analyses of the price elasticity of demand by income group can only be undertaken with cross-sectional (or panel) data. Governments commission and fund sample surveys to collect policy data on households and individuals. Donor organisations and market research bodies also collect data through surveys. These may be cross sectional (once-off, or regular but not directly comparable surveys) or panel surveys (which trace households or a cohort of individuals over time). Both can be valuable sources of tobacco policy information.

Data held by DATA Project PIs: Our researchers have negotiated access to tobacco data from several sources, for their own research. Researchers on our DATA Project are also involved in collecting tobacco data, as a means of filling in the tobacco data gaps on project countries. For example, one such project involves collecting retail prices of cigarettes in several African countries. We collected information on these datasets through a survey administered to project PIs. The survey questionnaire is *Appendix B* of this report. We also elicited information during meetings with project researchers, and from examining the data files and documents.

3. Data Scope (subject areas covered)

Time series data include variables like aggregate tobacco consumption (often broken down by type of tobacco), average prices, smoking rates (often by gender and age group), tax rates, tax structures, and tax revenue, over time.

Sample surveys measure tobacco use, such as the World Health Organisation's Global Adult Tobacco Survey (GATS). Health surveys can also be a source of tobacco data. Examples are the Demographic and Health Surveys of the DHS Program, which collects data on cigarette smoking and the use of other types of tobacco. The DHS covers 46 African countries. Income and expenditure surveys also gather data on tobacco-related expenditure and consumption.

4. Data Coverage

The project plans to collect and publish data from South Africa, and other African countries. The latter were initially Botswana, Kenya, Nigeria, and Senegal. We chose the countries based our experience with gaining permissions to use and share data from African sources. However, we were offered data from other countries and ended up being pragmatic and not rejecting any data we could obtain from other African countries. For example, our African cigarette price dataset covers Botswana, Lesotho, Mauritius, Namibia, South Africa, Swaziland, and Zimbabwe.

Data Management Plan

We also drew up a research data management plan (RDMP) for the Project, after discussions on this at the workshop. Research funders increasingly require research data management plans from researchers before they will commit funds to a project. Such plans provide information on how researchers will collect and manage data during the life of a project, and whether and how they will publish data for reuse once the project is completed. Several data management planning tools are available online, to assist research projects with this process. We used the DMPonline tool from the Digital Curation Centre to create our plan. As our project is data-focused, this was an essential component of our work. We use our RDMP as a guide to handling data we are collecting. Our project's RDMP is *Appendix C* of this report.

Our data management plan includes information about how we will source tobacco data, prepare them for reuse by tobacco researchers, and publish them online. We aim to manage the data per established protocols, and lessons learned from 15 years of handling research data. A model depicting how DataFirst manages data is available on their website. Data management includes the following steps:

- Sourcing data, and accepting data deposits
- Accepting or locating documents related to the collection and collation of these data
- Assuring data (checking data quality and confidentiality, anonymising data)
- Converting data to research-ready formats
- Storing preservation copies of the final dataset
- Preparing metadata (descriptions of the datasets)
- Uploading data and metadata to a branded page on our data publishing platform

Assisting IDRC researchers to monitor and report on our RDM planning involved giving feedback to the Project PI on project challenges and successes. Some of the IDRC Open Data case studies were given the opportunity to present on their Projects at SciDataCon 2016 in Denver, Colorado, USA, in September 2016. This was an opportunity to share experiences with and gain insight from data experts from around the world. Our session was organised by the IDRC and titled "Data sharing in a development context: The experience of the IDRC Data Sharing Pilot". Our presentation, *Collecting tobacco data from across Africa: The challenges of data aggregation and sharing* focuses on three datasets as case studies of the issues around opening up data from disparate sources. Our conference presentation is *Appendix E* of this report.

Project Outputs

Project outputs include our Project's data inventory (*Appendix A*), our Data Management Plan (*Appendix C*). A timeline of project activities is included in this report as *Appendix E*). Our presentation at SciDataCon 2016 is also a project output (*Appendix E*). As an output from the DATA Project, and as part of our Data Management Planning, we have created a branded page for DATA project output on our online data dissemination platform. The project page is highlighted on our portal's homepage.

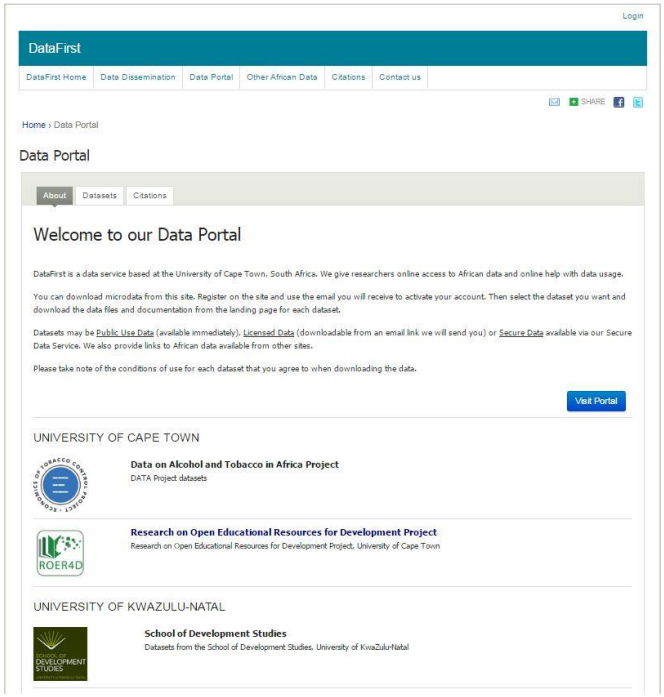

Figure 1: The DATA Project's branding on our data portal homepage.

Access to data from the project site will enable researchers to advise African governments on tobacco control policies in their countries. We have also begun to alert existing partner organisations working on tobacco research to data sources as they become available. Researchers who use our tobacco data will get ongoing help with their analyses via DataFirst's online user support site. Data on the site include:

- Data described on the site, but available elsewhere – we provide links to these data
- Data produced by other organisations, but which the Project has permission to share as Open Data
- Data collected by our project, which we share as Open Data.

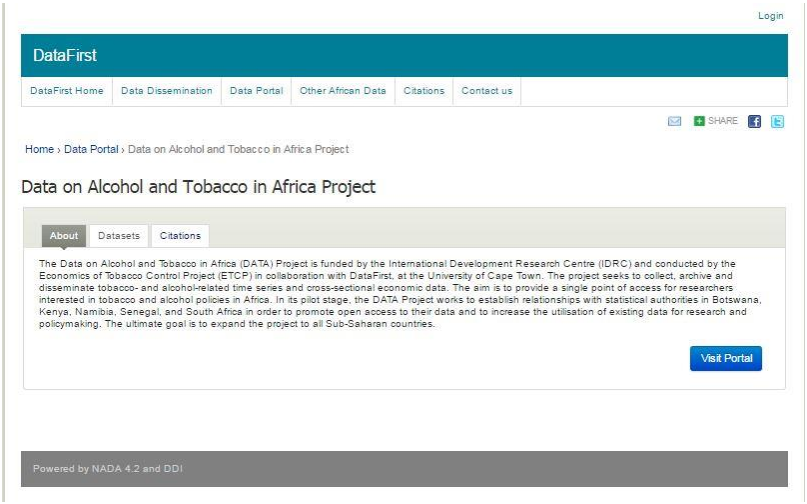

Figure 2: The DATA Project's webpage.

Some examples of our datasets are listed below:

African Cigarette Price Data 2015-2016

The DATA project organised University of Cape Town students to collect data on retail prices of cigarettes in their respective countries, using cellular phones. The students visited retailers and street vendors which sold cigarettes, and obtained permission from retailers to take photographs showing cigarette prices. They captured this data in excel spreadsheets. Working from these and the photos, project researchers and DataFirst's analysts cleaned the data and prepared a research-ready dataset. This data is now publicly available from a dataset landing page on our data portal

African subsets of international tobacco survey data

The DATA project is set up to source and share alcohol and tobacco data from African countries. However, as already noted, there are few tools to help researchers find out what data exists, and whether it is possible to obtain these data for secondary analysis. This situation leads to researchers wasting research time trying to locate data. It also leads to duplication of survey research in some countries, wasting limited research funds. The project has identified high-quality, easily obtainable African tobacco datasets from around the world and describe these datasets on the DATA project site. Metadata and documents on the datasets can be downloaded from the site in pdf format. We provide links to the sites from where the data can be downloaded. This should help researchers to discover existing tobacco data from African countries.

Problems and Challenges

The IDRC project hoped to gain insight on Open Data challenges from the case studies. Our main challenge was obtaining permissions to share the data more widely. We began by investigating accessibility of survey data for project countries. We started with the low-hanging fruit of WHO

survey data. However, ownership issues meant we could not share tobacco-related data from the WHO and other UN organisations. These data are already online, but they do not come with explanatory metadata. We are therefore aiding discovery of this through publishing metadata on and links to each dataset, via our clearinghouse. However, it would benefit African researchers if we could obtain permission to share the data directly from our site. Firstly, we could reformat the data to be optimally usable. We could also provide ongoing data usage support via our helpdesk.

We also attempted to obtain data from African National Statistics Agencies (NSAs). These agencies are the main collectors of primary data in Africa. We found links to African NSA data portals on the website of the International Household Survey Network (IHSN). The IHSN is a World Bank project which builds data curation and data publishing capacities of NSAs in developing countries. To this end, they have provided African NSAs with data management and dissemination tools and training. DataFirst has worked on this project, with the IHSN's Accelerated Data Program (ADP). Several African NSAs have built data publishing portals with the help of the ADP. We investigated tobacco data sources on these portals. The NSAs publish their data with "use and don't pass on" license requirements. Thus, we will not be able to share these as Open Data. We also applied for access for Project researchers. We were turned down for access to some of these datasets.

We also encountered technical difficulties when trying to register on the NSA websites (registration is required to download data). Research undertaken in 2009 and 2012 shows that many African NSAs encounter technical constraints when they try to curate and publish their data products (Woolfrey, 2013). This is largely due to a shortage of IT infrastructure and skills in these organisations. For example, the links to the data portal of the Nigerian Bureau of Statistics (NBS) were broken. We contacted the NBS who said they were working on the problem, but the site remains inaccessible.

Initially we assumed we could share data held by Principle Investigators on the DATA Project. However, it turned out that the data usage permissions obtained by PIs were strictly for their individual projects. We hoped to extend these permissions at least to other African researchers. However, data producers have not granted permissions for wider usage. This has thwarted some of our aims. However, it is important to note that this is a common occurrence, and needs to be investigated by research funders. Datasets 1 and 2, discussed below, are examples of common obstacles to opening research data. The first is that government data collectors often only share their data with a select number of researchers, on a case-by-case basis. The second is the commoditisation of public-domain data by "data brokers".

DATASET 1: Statistics South Africa's Cigarette and Alcohol Price data 2014-2016

This is an example of permissions only being granted to selected researchers. African governments collect data through surveys and administrative processes, and want to use these data as evidence for policy. However, they are reluctant to share these data formally with a wide academic audience. This could be because they do not have the infrastructure or technical knowledge to share securely, or because they do not have confidence in the quality of their data. It also stems from the bureaucratic nature of government departments (Woolfrey, 2013). Government data collectors therefore balance the need for expert policy input and their fear of exposure by sharing detailed data with a select group of trusted researchers. These investigators, in turn, are either happy to have exclusive access to official data, or do not want to jeopardise their access by insisting on more equitable sharing.

Statistics South Africa undertakes regular price surveys around the country, to calculate South Africa's Consumer Price Index. One of our DATA Project investigators obtained this price data for cigarettes and alcoholic beverages for his research. The data he holds spans 2014-2016, and includes

province, region, survey period, and prices by cigarette and alcoholic beverage brand. Statistics SA did not agree to us sharing this data, however. This is even though DataFirst has permission to distribute data from this agency. Statistics SA is well funded, and much of their data is also already in the public domain. It is therefore more likely that quality concerns and bureaucratic hurdles are reasons for their refusal. We will reapply for permissions to share these data in year two of the project.

DATASET 2: Tobacco Industry Data

This data is held by a researcher at our institution, and includes data on tobacco products manufacturing, tobacco prices, tobacco usage, illicit tobacco trade, and tobacco taxes, for Botswana, Egypt, Gambia, Kenya, Mauritius, Senegal, and South Africa. These data are very useful for tobacco control policy research. However, the data was purchased from a commercial data broker, and the commercial contract does not permit sharing. These data can therefore not be shared with other researchers.

However, the reuse of data for projects other than the original project purpose is common in academia, and investigators are often unaware that they are not complying with research ethics in these circumstances. We have seen grant proposals based on research using this type of unapproved data. This raises several ethical issues. Firstly, strictly speaking, researchers may not have consent to use the data for a different research purpose. Understandably, they will be reluctant to seek formal access, for fear of losing access to the data. Secondly, this leads to problems of research reproducibility. It is also difficult to confirm the quality of commercially produced datasets, as they cannot be scrutinised by the wider academic community. Lastly, much of this data is in the public domain, in company reports and government records. The Project will try and recreate this dataset from the public domain information. "Data brokers" often repackage and sell data which may be time-consuming to source and collate. Research funders could combat such commoditisation of public data by funding such unglamorous but vital data collation, data re-creation, and data rescue projects in LMICs.

Administrative Reflections and Recommendations

Our recommendations to the IDRC relate to opportunities to advance data sharing, highlighted by our project. Researchers and government institutions rely on external funding for data collection activities. Funders of research projects in developing countries are therefore in a strong position to promote Open Data practices in these countries. Funding organisations in several countries have linked research funding to commitment to data sharing on the part of funding recipients. The IDRC is gathering data on constraints to data access at the project level. In developing countries, government agencies are responsible for most data collection activities. IDRC investigations could include gathering data on obstacles for research use of government collected data in these countries. Perhaps project funding could be allocated for project participants to visit National Statistics Agencies in their countries to gather information on Open Data attitudes, infrastructure, and polices. As stated earlier, funding data collation of similar projects in developing countries could also make existing data easier to discover and use. This could benefit data-intensive research in these countries.

List of Sources Cited

Digital Curation Centre. DMPonline. Accessed 20160813 from https://dmponline.dcc.ac.uk/

G8 Open Data charter and technical index. 2013. UK Government Cabinet Office. http://opendatacharter.net/resource/g8-open-data-charter/

Statistics Canada quality guidelines. 2009. 5th edition. Ottawa: Statistics Canada (Catalogue no. 12-539-X). Accessed 20160812 from http://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-539-x2009001-eng.pdf

Woolfrey, Lynn. 2013. Leveraging data in African countries: Curating government microdata for research. Cape Town: DataFirst. (DataFirst Technical Paper Number 22). Accessed 20160821 from https://www.datafirst.uct.ac.za/images/docs/DataFirst-TP13_22.pdf

World Health Organisation. 2003. WHO framework convention on tobacco control. Geneva: WHO. Accessed 20170125 from http://apps.who.int/iris/bitstream/10665/42811/1/9241591013.pdf?ua=1

World Health Organisation. WHO fact sheet, online, accessed 20160809 from http://www.who.int/mediacentre/factsheets/fs339/en/

APPENDICES

APPENDIX A: Tobacco Data Inventory

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
Project 108098						
Initial audit of country tobacco data available	Administrative records of government departments		Data collection instruments	MSWord, pdf, excel spreadsheets	Small (Kilobytes, Megabytes)	Data discovery issues:
This will be to determine project countries			(administrative forms)	Access databases, excel spreadsheets	Medium to large (Megabytes, Gigabytes-Petabytes)	Very little data is available online
4 African countries, including South Africa	Department of Agriculture	Tobacco production data	Databases	MSWord, pdf	Small (Kilobytes, Megabytes)	Necessity of visiting countries to obtain data
	Department of Health	Prevalence of tobacco-related diseases	Reports			Data locked away in inaccessible formats
	National Statistics Agencies (NSAs)	Tobacco-related morbidity and mortality				e.g. non-digital, pdf
	Revenue Service	Tobacco taxation data				Departments do not know what data they have, or where they are located
	Department of Trade and Industry	Tobacco products manufacturing data				
		Tobacco import/export data				Big Data Issues - storage and throughput
						Administrative database could

Formatted: French (Canada)

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
						grow to many petabytes
	Census and sample survey data	Tobacco production data	Data collection instruments	MSWord, pdf		
		Prevalence of tobacco-related diseases	(questionnaires, diaries)			Data discovery issues:
	Department of Agriculture	Tobacco-related morbidity and mortality	Census/survey reports	MSWord, pdf. Online		Very little of this data is listed online
	Department of Health	Cross-country data	Census and survey datasets	Access, .csv, SPSS, Stata,		Few provide good metadata for discovery
	National Statistics Agencies (NSAs)		(data files, documents, programs)			Need to work with International Donor Org to source data
	Universities and Research Institutes					
	International Donor Organisations (IDOs)					
	Tobacco industry records	Cost of tobacco production				
		Prices of raw tobacco				
		Price per cigarette pack (most popular brand/imported brands/all)				
		Average price per standard cigarette pack (20 cigarettes) (total revenue and total sales)				
		Prices of other tobacco products (by kind/brand name/unit)				
		Salaries in tobacco industry				

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
		Capital investment in tobacco industry				
		Advertising spend of tobacco industry				
		Profit of tobacco industry - domestic firms				
		Profit of tobacco industry - importing firms				
		Income tax paid by tobacco industry				
		Wholesale, retail margin of tobacco industry				
		Market share by tobacco company				
		Market share by brand/ most sold brands				
		Production of tobacco products				
		Sales of tobacco products				
		Mergers and acquisitions of tobacco companies				
		Foreign investment and (including regulation)				
Obtaining data from Project countries	Administrative records of government departments		Data collection instruments	MSWord, pdf, excel spreadsheets	Small (Kilobytes, Megabytes)	Data access issues:
			(administrative forms)	Access databases, excel spreadsheets	Medium to large (Megabytes, Gigabytes-Petabytes)	Problems around negotiating permissions to access and use data
	Department of Agriculture	Tobacco production data	Databases	MSWord, pdf	Small (Kilobytes, Megabytes)	May have to negotiate this for each project

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
	Department of Agriculture/Tobacco farming sources	Cultivation area of tobacco crops				
		Total tobacco production				
		Tobacco farm employment				
		Farm gate prices of tobacco				
		Import prices of tobacco				
		Export prices of tobacco				
	Department of Health	Prevalence of tobacco-related diseases	Reports			No tradition of sharing admin data for research
		Disease burden - annual deaths attributable to tobacco use				
		Exposure to second-hand smoking (% children exposed)				
		Tobacco use (cigarettes/other tobacco products/all)				
		Legal environment - regulations on health warnings, smoke-free areas				
	National Statistics Agencies (NSAs)	Tobacco-related morbidity and mortality				Govt databases not inter-operable (in-house/proprietary)
	Revenue Service	Tobacco taxation data				
	Department of Trade and Industry	Tobacco products manufacturing data				

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
		Tobacco leaves - import/export data (volume)				
		Tobacco leaves - import/export data (volume)				
		Tobacco products - import/export data (volume)				
		Tobacco products - import/export data (volume)				
		Major partners in tobacco trace				
	Internal Revenue Service/Other government sources	Tobacco tax data				
		Tax structure/rate (historical)				
		Tax base				
		Tax revenue by type of tax				
		Total tax revenue				
		Average excise tax per cigarette pack (total tobacco excise revenue/total sales)				
		Average total tax per cigarette pack (total tobacco tax revenue/total sales)				
		Average excise tax per other unit of tobacco (total tobacco excise revenue/total sales)				
		Average total tax per other unit of tobacco				

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
		(total tobacco tax revenue/total sales)				
	Census and sample survey data	Tobacco production data	Data collection instruments	MSWord, pdf		Data access issues:
		Consumer Price Index (CPI) for cigarettes				
		Prevalence of tobacco-related diseases	(questionnaires, diaries)			Negotiating project (and ongoing) access
	Department of Agriculture	Tobacco-related morbidity and mortality	Census/survey reports	MSWord, pdf. Online		Most African census data is not in the public/research domain
	Department of Health	Household expenditure on tobacco products (cigarettes/other)	Census and survey datasets	Access, .csv, SPSS, Stata,		Only some African governments share their sample survey data
	Dept of Health/ other Depts	Tobacco use - average age of initiation				
	National Statistics Agencies (NSAs)	Tobacco use - consumption intensity (e.g. no. cigarettes per day)	(data files, documents, programs)			Will rely heavily on survey data from IDOs
	Universities and Research Institutes	Cessation - share of former smokers				
	International Donor Organisations (IDOs)	Cessation - % of smokers with attempts to quit (no. of attempts)				
		Cessation - relapse rate				
		Cross-country data				

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
						Need to draw on contacts at African NSAs
Quality controlling, preserving and sharing data	Administrative records of government department		Data collection instruments	MSWord, pdf, excel spreadsheets	Small (Kilobytes, Megabytes)	Data quality issues:
			(administrative forms)	Access databases, excel spreadsheets	Medium to large (Megabytes, Gigabytes-Petabytes)	Accuracy - issues around data collection
	Department of Agriculture	Tobacco production data	Databases	MSWord, pdf	Small (Kilobytes, Megabytes)	Reliability - trustworthy sources? Sources contradict each other
	Department of Health	Prevalence of tobacco-related diseases	Reports			Comparability - over time periods and between domains
	National Statistics Agencies (NSAs)	Tobacco-related morbidity and mortality				Timeliness - time lags in collection/collation of data
	Revenue Service	Tobacco taxation data				Interpretability - almost no data documentation, metadata
	Department of Trade and Industry	Tobacco products manufacturing data				
		Tobacco import/export data				Skills and infrastructure exist for this at DataFirst
						Big Data issues - storage and throughput
						Administrative database could

Process/Work Package	Data Source	Data Description	Data Type	File Format	File Size	Potential Issues
						grow to many petabytes
						Need computing power for data linking e.g. admin and panel data
	Census and sample survey data	Tobacco production data	Data collection instruments	MSWord, pdf	Small (Kilobytes, Megabytes)	Data quality issues:
		Prevalence of tobacco-related diseases	(questionnaires, diaries)		Medium to large (Megabyes, Gigabytes-Petabytes)	Accuracy - issues around chances in data collection methods
	Department of Agriculture	Tobacco-related morbidity and mortality	Census/survey reports	MSWord, pdf. Online	Small (Kilobytes, Megabytes)	Reliability - issues with data collection, collation methods
	Department of Health	Cross-country data	Census and survey datasets	Access, .csv, SPSS, Stata,		Comparability - widely different findings among surveys
	National Statistics Agencies (NSAs)		(data files, documents, programs)			Timeliness - time lags in collection/collation of data
	Universities and Research Institutes					Interpretability - limited data documentation, metadata
	International Donor Organisations (IDOs)					

APPENDIX B: Questionnaire for survey of data held by DATA Project researchers

SURVEY OF DATASETS HELD BY RESEARCHERS IN THE ECONOMICS OF TOBACCO CONTROL PROJECT, UNIVERSITY OF CAPE TOWN

Background and purpose

The Economics of Tobacco Control Project (ETCP) provides technical assistance to policy makers and tobacco control stakeholders in a number of African countries to increase research capacity in the economics of tobacco control in Africa. The ETCP was approached by the Secretariat of the WHO Framework Convention on Tobacco Control (FCTC) to become a Knowledge Hub supporting Article 6 (tobacco taxation) and Article 15 (illicit trade) of the FCTC. The Knowledge Hub at the ETCP intends to help individual countries with tax modelling (e.g. predicting the likely impact of a change in the excise tax structure or the excise tax rate on cigarette consumption, smoking prevalence and government revenue), other aspects of tobacco tax policies, and with estimating the size of illicit trade, for example.

Such research requires good data. The ETCP has been funded to establish a clearinghouse for tobacco economic and policy data in sub-Saharan Africa. Preliminary work includes an audit of existing data sources held by ETCP researchers which may be suitable for this clearinghouse. The Tobacco Clearinghouse Project has an explicit Open Data agenda. It is therefore important to discover which data held by the project is suitable for inclusion as Open Data.

Instructions

Please take time to complete the questionnaire to support the Project. Your participation in this questionnaire is voluntary. It should take about 5 to 10 minutes to complete and the questions are mostly multiple choice.

Section 1: Personal details

This section asks you questions about your position and your research project or research interest.

1.2 Please indicate your Position at the University of Cape Town.

a. Postgraduate research student
b. Academic/research staff
c. ECTP Principle Investigator
d. Research administrator
e. Other – please specify

1.3 Please provide a description of your research area of interest

Section 2: Digital research data held by you for your research

This section asks about the research data held by you for the ETCP project, including about data sources, data ownership, data formats, and size and importance of the data.

Please circle the most appropriate answer(s)

2.1 Please indicate the data source. (Please select all that apply)

a. Raw data generated by a program
b. Raw data from instruments (including questionnaires)
c. Laboratory notes
d. Patient data
e. Data from qualitative research
f. Other – please specify

2.2 Who owns the tobacco data held by you?

2.3. Please select the data formats that apply to your research or project

a. Data in a database (e.g. MySQL, Oracle)
b. Images, scans or x-rays
c. Digital audio
d. Digital video
e. Text documents (e.g. Word, PDF)
f. Spread-sheets (e.g. Excel)
g. Raw data in American Standard Code for Information Interchange (ASCII)
h. Raw data is a proprietary software format such as SAS/SPSS/Stata
i. Geospatial data - vector (e.g. co-ordinate lists, CAD files, shape files, geo-databases)
j. Geospatial data - raster (e.g. scanned maps, satellite imagery, aerial photography)
k. Other – please specify

2.4 Please indicate the size of your data holdings in Kilobytes/Megabytes/Gigabytes

2.5 Please indicate the importance of the data.

Vital (i.e. you could not continue your research if you lost the data)
Important
Ephemeral

---

3. Management of the Data

This section asks about the management of the data you hold for the ETCP

3.1 Who is currently responsible for managing the data? (Please select all that apply.)

a. Project manager
b. Designated person on project
c. External project partners

d. IT staff within the department
e. UCT IT staff
f. Research assistant
g. Yourself
h. National data archive
i. Discipline-specific national archive
j. Discipline-specific international archive
k. Nobody
l. Don't know
m. Other - please specify

3.2 Retention period: How long you aim to keep the data?

a. Only over the project period
b. Up to 5 years
c. Up to 10 years
d. More than 10 years
e. Don't know

3.3 Where are the data stored?

a. UCT ICTS server
b. Departmental server
c. CDs/DVDs
d. USB/Flash drives
e. External hard drives
f. Tapes
g. Third party/ Cloud/ Commercial data service
h. Don't know
i. Other - please specify

3.3 How frequently have the data been updated since you began your research?

a. Never
b. Daily
c. Weekly
d. Monthly
e. Annually
f. Don't know

3.4 Are your data holdings backed up regularly?

a. Yes
b. No
c. Don't know

3.5 If “Yes” to question 12, where are the data holdings backed up? (Select all that apply)

j. UCT ICTS server
k. Departmental server

l. CDs/DVDs
m. USB/Flash drives
n. External hard drives
o. Tapes
p. Third party/ Cloud/ Commercial data service
q. Don't know
r. Other - please specify

3.6 Do you currently have a formal Research Data Management Plan in place for the data you hold for the ETCP?

a. Yes
b. No
c. Don't know

---

4. Data sharing
This section deals with sharing of data you hold for the ETCP

4.1 Do you have a funding requirement to share your data?
a. Yes
b. No
c. Don't know

4.2 Do you currently allow others to access the data you use?

a. Yes
b. No
c. Don't know

4.2 If yes, who has access to your data? (Please select all that apply.)

a. Project partners
b. Students / colleagues in your department
c. Any researcher who requests the data
d. General public
e. Only to confirm published research from the project
f. Other - please specify

4.3 If no, what issues related to access prevent you from sharing the data? (Please select all that apply.)
a. Requirement to publish extensively before sharing
b. Confidentiality
c. Intellectual Property Rights
d. Commercial nature of the data (data is for sale from a third party)
e. Possible misinterpretation of data
f. Time/effort required preparing data for sharing
g. Other - please specify

5. Support and services

This section deals with support and services around the data you hold for the ETCP.

5. 1 Have you ever encountered the following data-related problems

a. Inadequate shared space for data storage
b. Loss of data files
c. Difficulties with finding data because of lack of standard file names / formats
d. Inability to use files in older formats
e. Confusion with colleagues around versions of the data

5.2 What support would you like from the Tobacco Data Clearinghouse?

a. Secure storage facilities for data
b. Access to data from other African tobacco related projects
c. Access to data from the tobacco industry
d. Access to tobacco data from African government sources
e. Data curation (management and preparation for reuse) services
f. Guidance on using your data

6. Comments and thank you

If you have any comments about this questionnaire, or would like to provide more information about the tobacco data you hold for the Project or your research, please use the space below.

Thank you for your time and willingness to complete the questionnaire

Acknowledgements to authors of the following example questionnaires:

- Edinburgh Data Audit Implementation: Online Questionnaire
- Imperial College London DAF Survey Questionnaire
- University of Southampton Questionnaire
- University of Oxford Interview Framework
- University of Glasgow Digital Preservation Study: Interview Template

Listed in: University of Edinburgh. 2009. Data Asset Framework implementation guide, accessed 20130618 from http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf

APPENDIX C: DATA Project: DATA MANAGEMENT PLAN

1. ADMINISTRATIVE DETAILS

Project Name: Opening access to economic data to prevent tobacco related diseases in Africa

Project Identifier: PROP0715E

Grant Title: 108098-001

Principal Investigator / Researcher: Hana Ross

Project Data Contact: Lynn Woolfrey, +27216505707, lynn.woolfrey@uct.ac.za

Description: The purpose of this project is to demonstrate that data for tobacco control research in sub-Saharan Africa can be collected and distributed from an Open Data platform and used in policy relevant research activities. The platform and data will improve the capacity for tobacco control research in key sub-Saharan African countries, and help develop a continent-wide research approach to tobacco control.

2. DATA COLLECTION

What data will you collect or create?

See our data inventory which is Appendix A of this document

How will the data be collected or created?

Data Collection methods:

1. Desk-based search of official websites of project countries:

This will involve searching websites of government departments of selected Sub-Saharan African countries for administrative data collected by these departments which relate to tobacco production or usage.

For example, these websites will be scraped for data on:

a. Tobacco production data (Departments of Agriculture)

b. Prevalence of tobacco-related diseases, and tobacco-related morbidity and mortality (Departments of Health)

c. Tobacco taxation (Internal Revenue Services)

d. Tobacco products manufacturing, tobacco imports/exports (Departments of Trade and Industry)

National Statistics Agencies (NSAs) websites are another useful place to find administrative data. In South Africa unit record administrative data from departments, repackaged as research datasets, are shared by the NSA

If data collection instruments (administrative forms) used to collect the data are available on these sites they can provide useful information on data fields and subject categories for final datasets

We will follow this up with a study of useful variables on tobacco from country surveys.

Again, an initial desk-based study will allow us to download public use data from project countries, and interrogate these for tobacco-related variables. From this a "question bank" will be created of useful variables and the datasets where these can be found.

2. Desk-based search of industry websites

The second component of our desk-based research will involve examining online records of the tobacco industry. From these we hope to obtain data on:

Cost of tobacco production, and profits, in the industry, prices of raw tobacco and tobacco products, salaries, capital and foreign investment, mergers and acquisitions, advertising spend, and regulations in the industry

3. Approaches to data holders

Our desk search may reveal the existence of datasets with a tobacco data component but which are not in the public domain. In these cases, we will approach NSAs or the relevant research projects in the project countries to release this data and allow the Project to host this on their Open Data portal. This may be a fraught process but any challenges and successes can be written up to inform our future work.

4. Own surveys

The project has already crowd-sourced data on current prices of tobacco products in two project countries. This may be expanded during the project to all project countries.

3. DOCUMENTATION AND METADATA

What documentation and metadata will accompany the data?

Supporting documents

These documents will be shared with the data files, where available. Forms used for collecting administrative data will be shared with administrative datasets. Data collection instruments (questionnaires, diaries) will be made available with the survey data. Code lists used in collecting the data will also be provided. Final reports from data collection projects will form part of datasets, where available.

Metadata

Each dataset will have a metadata record to help data users analyse the data. This metadata record will be created during examination of the data and data collection instruments. It will include information gathered on the dataset during the data collection process. The latter is often useful for those analysing the data. Issues around data quality will form part of the metadata record. Metadata will be created according to the Data Documentation Initiative(DDI) international metadata standard, using Nesstar Publisher, which is free data markup software for the creation of XML compliant metadata according to the DDI standard.

4. ETHICS AND LEGAL COMPLIANCE

How will you manage any ethical issues?

The administrative data we will collect will mostly be in the public domain, in the form of reports and other records from government departments. The survey data we will collect will be anonymised data already shared with researchers, although not always online. The industry data will be data made available to shareholders and the public. We are adding value by bringing these sources together and providing a means for researchers to easily discover and download these data.

However, we will endeavour to make data available that is not yet in the public domain. In these cases, we will ensure that:

a. We have the necessary permissions from data owners make these data open.

b. The data is suitably anonymised, to protect respondent confidentiality and privacy

c. We take national laws on sharing data across borders into account. Where such restrictions exist, we will be unable to host this data.

d. We work with all stakeholders to ensure agreement on what will be shared, how, and with whom.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The government data we will collect is not subject to IPR. The tobacco industry data we will collect will be data made public by the industry, which does not publish information that would compromise their IP rights. However, we will check each tranche of data we obtain, to ensure we have permission to pass the data on to third parties.

5. STORAGE AND BACKUP

How will the data be stored and backed up during the research?

The data would be stored on a server managed and backed up by the University of Cape Town's Commerce IT Department. Curation of the data will be the responsibility of DataFirst's Research Data Service. DataFirst is a technical partner on the Project. Each preservation dataset will consist of data files, document files, metadata files, and any programme files used in creating the data files. Data Service staff will be responsible for adding data updates to datasets. We will also handle version control to ensure the most recent and accurate data files are published, and earlier versions are available for verification or replication of research which may cite earlier versions of the data.

How will you manage access and security?

Access to the server hosting the preservation datasets will be password controlled. Passwords will be allocated by the Commerce IT manager only to Data Service staff. Server software will monitor data security and integrity.

6. SELECTION AND PRESERVATION

Which data are of long-term value and should be retained, shared, and/or preserved?

Criteria for preservation will be:

a. Data is tobacco-related

b. Data covers project countries

c. Data is accurate and reliable (we will undertake quality audits to determine this)

d. Data is unit record data not available in another dataset in the collection

e. Data is not readily available from another repository

Retention: It is difficult to predict what data has long-term value. Our policy will be to store unit record tobacco data indefinitely. As these datasets grow so will their value over time. Time series data continue to be useful for economic and health policy research in the long term.

Sharing:

Because we aim to establish an Open Data portal, all data retained/preserved will also be shared. The Project's policy is aligned to DataFirst's policy: We do not archive data which cannot be shared with researchers in some form and at some access level.

What is the long-term preservation plan for the dataset?

There will be numerous datasets. Our long-term preservation plan for the Project's data holdings depends on the sustainability of DataFirst's Research Data Service. The service was established in 2001 and is a unit at the University of Cape Town, a well-funded and well-established university in South Africa. Our sustainability prospects are therefore good.

7. DATA SHARING

How will you share the data?

The data will be shared as discrete datasets (by country, year, data source). DataFirst hosts and shares data via an online dissemination platform, based on the National Data Archive Open Source software developed by the World Bank's Development Data Group

The platform provides several data access options. The Project's data will be shared as Public Use data. That is, researchers will need to register on our site and say for what purpose they will use the data, but access will be immediate and automatic, with no vetting of use. The usage information we collect will be to support service improvements.

The data will be shared in several formats, including excel spreadsheet, and data files in the commonly used statistical analysis programmes (SPSS, Stata). We will also make the data available as .csv files, in line with Open Data requirements.

Are any restrictions on data sharing required?

We aim to share the tobacco data we collect as public access data. We do not aim to support research-use only requirements, as this is counter to Open Data Principles. Policy research, academic research, business analysis, and private sector innovation all need good data, and countries benefit from informed decision-making in all these spheres.

8. RESPONSIBILITIES AND RESOURCES

Who will be responsible for data management?

The Manager of DataFirst's Research Data Service will be responsible for curating the Project's data. This is in line with the project proposal for DataFirst to be funded to provide technical support. DataFirst's Manager has 25 years' experience in managing research data and working with data users.

Experience from undertaking data rescue projects in South Africa will also be useful in assisting with data collection activities.

What resources will you require to deliver your plan?

Funding for data collection has been budgeted for in the Project. This may need to include funding to travel to project countries and negotiate with data collectors in government and academia to release their data, and allow its reuse. Funding has been provided for a Project Manager. The Project Manager is responsible for conducting online data audits and downloading data, and populating the database which DataFirst's Manager will curate. This will be a time- and labour- intensive task and more staff hours may need funded for this.

APPENDIX D: DATA PROJECT: TIMELINE OF ACTIONS

20160309 – We attend an introductory workshop for project participants from 9-10 March 2016 at the IDRC in Ottawa, Canada, where we begin an inventory and data management plan for the project.

20160406 – We finalise an inventory of tobacco data sources and types targeted by the project

20160429 – We complete a Research Data Management Plan (RDMP) for the project, using the DMPonline template, to provide an overview of how we will curate and share project data.

20160714 – The ETCP PI deposits secondary cigarette price data by brand from 2014-2016 obtained from Statistics South Africa (StatsSA).

20160525– We begin an audit of survey data available from National Statistics Agencies in Project countries.

20160517 – We make direct contact with the research group, Consortium pour la recherche Economique et Sociale (CRES) in Senegal regarding data availability and data sharing.

20160629 - We register on African NSA data sites, and apply for datasets for the ETCP. We also download the metadata records for all tobacco-related datasets

20160801 – Project Coordinator obtains data from project investigators and passes these on to DataFirst

20160812 – We meet with the co-ordinator of the African tobacco retail price survey project to complete a metadata record for this data

20160816 – Survey on research data management of tobacco data held by Project team-members is undertaken – we send a questionnaire to each researcher on the team who holds tobacco data, to complete for their dataset(s)

20160831 – We prepare a template for metadata on datasets held by each project researcher, and send these to them for completing

20160901 – Project researchers complete metadata forms for their data

20160913- We present on our project in the IDRC's session "Data Sharing in a Development Context: The experience of the IDRC Data Sharing Pilot" at SciDataCon 2016 in Denver, Colorado, USA.

20161002- Project investigator confirms that Dataset 2 (African tobacco industry data) was purchased from a commercial third party, and cannot be shared

20161010 – Project investigator sends a request to Statistics SA for permission to share the South Africa CPI cigarette price data on our data site.

20161027 – Statistics SA turns down our request to share their cigarette price data.

20161107 – The Project portal goes live, initially with seven datasets.

20161201- We attend a wrap-up meeting for Project 108131 at the IDRC in Ottawa Canada, from 1-2 December, 2016. The workshop focused on lessons learned by project participants and Program Officers. It also covered the support needed by participants and Program Officers to enable data sharing from IDRC funded projects.

20170125 – Final project report submitted.

APPENDIX E: SCIDATACON 2016 PRESENTATION